Listening. Learning. Leading.®

Research Report
ETS RR–13-24

TOEFL11: A Corpus of Non-Native English

Daniel Blanchard

Joel Tetreault

Derrick Higgins

Aoife Cahill

Martin Chodorow

November 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

TOEFL11: A Corpus of Non-Native English

Daniel Blanchard
Educational Testing Service, Princeton, New Jersey

Joel Tetreault
Nuance Communications, Inc., Sunnyvale, California

Derrick Higgins and Aoife Cahill
Educational Testing Service, Princeton, New Jersey

Martin Chodorow
Hunter College and the Graduate Center of CUNY, New York, New York

November 2013

Action Editor: Beata Beigman Klebanov

Reviewer: Jill Burstein

Abstract

This report presents work on the development of a new corpus of non-native English writing. It will be useful for the task of native language identification, as well as grammatical error detection and correction, and automatic essay scoring. In this report, the corpus is described in detail.

Key words: native language identification, corpora, *TOEFL*® test

This report presents a new publicly available corpus of non-native English writing called TOEFL11.[1] TOEFL11 consists of essays written during a high-stakes college-entrance test, the *TOEFL*® test. The corpus contains 1,100 essays per language sampled as evenly as possible from eight prompts (i.e., topics) along with score levels (low/medium/high) for each essay.

1 Motivation

The release of the TOEFL11 corpus is intended to support a broad range of research studies in the fields of natural language processing (NLP) and corpus linguistics. The corpus was compiled with the specific task of NLI in mind, but it will likely be useful for tasks and studies in the educational domain as well.

1.1 Native Language Identification

In recent years, there has been growing interest in the NLP task of NLI, which aims to determine the native language (L1) of the author of a text (given some predetermined universe of possible L1s). As Tetreault, Blanchard, Cahill, and Chodorow (2012) indicated, the NLI task may be a useful first stage in author identification or may be used to tailor the sorts of feedback provided to developing writers in formative educational contexts. More generally, this task may be useful in any context for which a model of the user is to be developed, including security applications, targeting of advertising, and market research. Koppel, Schler, and Zigdon (2005) spurred the recent interest in this task in the NLP community with their study using the International Corpus of Learner English (ICLE). Other recent studies include those of Wong and Dras (2011), Wong, Dras, and Johnson (2011), Swanson and Charniak (2012), and Tetreault, Blanchard, et al. (2012). To our knowledge, all previous work on this task has focused on texts in English, although the task could be addressed with any second language (L2), in principle.

Almost all of the studies addressing the NLI task have used the ICLE described by Granger, Dagneaux, and Meunier (2009). This corpus contains 6,085 essays written by undergraduate university students who were non-native English speakers. This collection of essays provides a useful resource for exploring the characteristics of non-native English writing in general, but because it was not designed specifically for the NLI task, it has two characteristics that limit its suitability for NLI. The first is that the distribution of essay topics is not even across the various L1s in the ICLE. Because the essay topic drives the vocabulary usage in a student's essay, if

the topic distribution differs substantially by L1, the NLI task can become conflated with the identification of an essay's topic. The second limitation of the ICLE for NLI is that, because of differences in the way essay tasks were administered and responses collected, there are differences in character encodings and annotations across languages as well. These differences provide cues to the L1 of an essay that would not generalize to other contexts. See Tetreault, Blanchard, et al. (2012) for further discussion.

The TOEFL11 corpus was designed specifically to support the task of NLI. Because all of the essays were collected through ETS' operational test delivery system for the TOEFL test, the encoding and storage of all texts in the corpus is consistent. Furthermore, the sampling of essays was designed to ensure approximately equal representation of L1s across topics, insofar as this was possible (cf. Section 2.2). Finally, TOEFL11 is larger than the ICLE subset typically used for NLI, comprising a total of 12,100 essays, and contains more L1s (11, compared to seven for the ICLE subset).

1.2 Educational Applications of Natural Language Processing and Corpus Linguistics

More generally, multiple research communities have demonstrated growing interest in using natural language corpora to support educational applications. As the only corpus of its size containing non-native English writing samples from a standardized, meaningful, and authentic assessment context, TOEFL11 will certainly be a useful resource to support these research directions.

In the field of computational linguistics, a growing number of researchers have begun to work on educational problem spaces. The increase in the resources dedicated to educational work is reflected in the growth of education-focused NLP workshops such as the NAACL Workshop on the Innovative Use of NLP for Building Educational Applications (Tetreault, Burstein, & Leacock, 2012) now in its seventh installment, and the ISCA Speech and Language Technologies in Education Workshop (SLaTE),[2] hosted for the sixth time in 2013. The journal *Speech Communication* featured a special issue on educational applications in 2009, and the 2013 shared task for the Conference on Natural Language Learning (CoNLL) related to the educational task of automated grammatical error detection. Corpus linguists have also made use of existing texts from educational domains for some time (Biber et al., 2004), and the importance of educational tasks

to this field is attested by the specific request for papers related to the use of corpora in language teaching and learning in the call for papers for the 2013 meeting of the American Association for Corpus Linguistics.[3] Finally, there has been a recent increase in the attention devoted to educational applications of NLP in the fields of language testing and educational measurement, due to the increasing use of automated methods for scoring open-ended responses in large-scale standardized testing and the expectation of future expansion of this use. The annual meetings of both the National Council on Measurement in Education and the Language Testing Research Colloquium in 2012 featured multiple symposia exploring new technologies for automated analysis of spoken and/or textual test responses.

Previous work in these fields has analyzed other, publicly available sets of student essays. However, each of these existing sets of data differs in significant ways from the TOEFL11 corpus, which makes this new resource of significant incremental value. The ICLE corpus distributed by UC Louvain was discussed above in connection with the task of NLI. The TOEFL11 corpus may be better suited than ICLE for some other kinds of research studies as well, given its relatively even distribution of topics across L1 groups and the uniform administration conditions for the TOEFL11 essays.

Another publicly available corpus that has been used in previous work is the set of Cambridge First Certificate of English (FCE) exam scripts analyzed by Yannakoudakis, Briscoe, and Medlock (2011). This set of essays differs from TOEFL11 in a number of ways. First, it provides actual essay scores, rather than the score levels associated with the TOEFL11 essays. Second, the FCE dataset includes essays written by a total of only 1,244 examinees, (compared to the 12,100 examinees in TOEFL11). Third, because the FCE dataset includes many more essay topics than TOEFL11, it includes fewer essays written on each topic. Fourth, the FCE dataset contains manually annotated grammatical errors for each essay. Finally, the FCE dataset includes two essays from each writer, whereas TOEFL11 includes only a single essay for each student. TOEFL11's size may make it better suited for some studies, while the richer information per examinee that the FCE corpus provides may make it more suitable for others.

A large set of essays was also released recently through the Kaggle[4] platform for hosting competitive statistical modeling tasks. This set of 17,450 essays across eight prompts was released as part of the Automated Student Assessment Prize (ASAP) challenge sponsored in 2012 by the Hewlett Foundation (Shermis & Hamner, 2012), in an effort to gauge the

contribution that automated essay scoring systems could make to the design and execution of state assessment systems. The essay prompts and responses were ultimately provided by state education departments in the United States and are therefore drawn from a sample that accurately represents the state of writing in U.S. middle and secondary schools. The usefulness of this corpus for other research purposes is limited, however, by the format in which essay texts are provided. The ASAP essays were preprocessed by a routine intended to anonymize the text and to ensure that no sensitive information about test takers (which they may have chosen to include in their responses) was released to the public. This preprocessing was fairly aggressive and expunged both named entities and most other capitalized words, replacing them with special tags.

In addition to the corpora described above, there are many other datasets containing learner essays that have been used for the task of automatic grammatical error correction. Enumerating them is outside the scope of this report so we refer the reader to Leacock, Chodorow, Gamon, and Tetreault (2010). Some notable corpora include the Chinese Learners of English Corpora (CLEC)[5] and the National University of Singapore Corpus of Learner English (NUCLE).[6] CLEC is a collection of essays written by Chinese English language learners comprising one million words and includes error annotations. It has been used in the grammatical error correction work of Gamon et al. (2008) and Rozovskaya and Roth (2010). The NUCLE corpus contains 1,400 essays written by National University of Singapore students and also comprises one million words and is error-annotated. It has been used in the work of Dahlmeier and Ng (2011). The corpus was also used in the CoNLL 2013 Shared Task Competition on Grammatical Error Correction.[7]

Given these substantial differences between TOEFL11 and other extant corpora of student essays, we anticipate that this corpus will be widely used in studies on a variety of topics, including automated scoring of essays, automated detection of grammatical errors, corpus linguistic analyses of linguistic features across L1s, cross-genre comparisons of writing, and of course, NLI.

2 Corpus Description

2.1 *TOEFL iBT*® Essays

The 12,100 essays of the TOEFL11 corpus are responses provided by test takers to the *TOEFL iBT*® test in 2006–2007. The TOEFL test is used internationally as a measure of academic English proficiency, among other purposes, to inform admissions decisions for students seeking to study at institutions of higher learning where English is the language of instruction.

The test includes reading, writing, listening, and speaking sections and is delivered by computer in a secure test center. The TOEFL test (comprising all four sections) takes an average of 4 hours to complete.

The writing section of the test consists of two essay tasks that differ in structure. The *independent* task asks students to write an essay in response to a brief writing topic. Figure 1 illustrates a sample independent task, together with the on-screen instructions provided to students. The *integrated* task asks students to first read a short passage presenting one perspective on a topic and then listen to a short lecture, presenting a different perspective. The student is asked to summarize and synthesize these perspectives in an essay. All of the essays in the TOEFL11 corpus were taken from the TOEFL independent task.

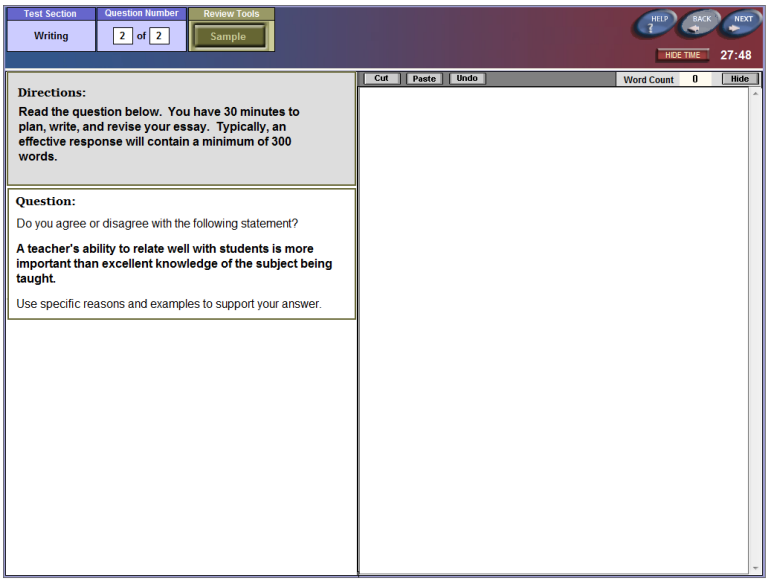

***Figure* 1 Screenshot of sample TOEFL iBT independent writing task.**

Independent writing task responses are scored on a 5-point scale, according to criteria described in the rubric available online.[8] Currently, each essay is scored twice: once by a trained human rater and once by the *e-rater*® engine for automated scoring of essays (Attali & Burstein, 2006). Because the data for the TOEFL11 corpus predates the usage of e-rater for scoring TOEFL essays, all scores for the essays were given by human raters.

2.2 Sampling

TOEFL11 was generated by sampling from a set of TOEFL essays written in 2006 and 2007 from eight retired prompts. As the goal was a corpus with a large number of essays per L1 that was evenly distributed across language and prompts, only L1s for which the set contained at least 1,100 essays were considered when sampling. There were 11 L1s that met this cut-off: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. Only essays for which the author provided permission for research use of their responses were selected. The number of essays from each L1 for each topic are shown in Table 1 and Figure 2. There is an average of 348 word tokens per essay in TOEFL11.

Table 1
Number of Essays per Language per Prompt

Language	P1	P2	P3	P4	P5	P6	P7	P8
Arabic	138	137	138	139	136	133	138	141
Chinese	140	141	126	140	134	141	139	139
French	158	160	87	156	160	68	151	160
German	155	154	157	151	150	28	152	153
Hindi	161	162	163	86	156	53	158	161
Italian	173	89	138	187	187	12	173	141
Japanese	116	142	140	138	138	142	141	143
Korean	140	133	136	128	137	142	141	143
Spanish	141	133	54	159	134	157	160	162
Telugu	165	166	167	55	169	41	166	171
Turkish	169	145	90	170	147	43	167	169
Total	1,656	1,562	1,396	1,509	16,48	960	1,686	1,683

2.3 Essay Scores

The score levels provided in the corpus were calculated first by combining the individual 5-point-scale scores given by the human raters and then by collapsing this combined score into a 3-point scale (low/medium/high). The 5-point-scale human scores were combined using the following rules:

1. Average the two scores on each item if the two scores differ by no more than 1 point (adjacent scores). For example, if the two scores are 3 and 3, the item score is 3. If two scores are 4 and 5, the item score is 4.5.

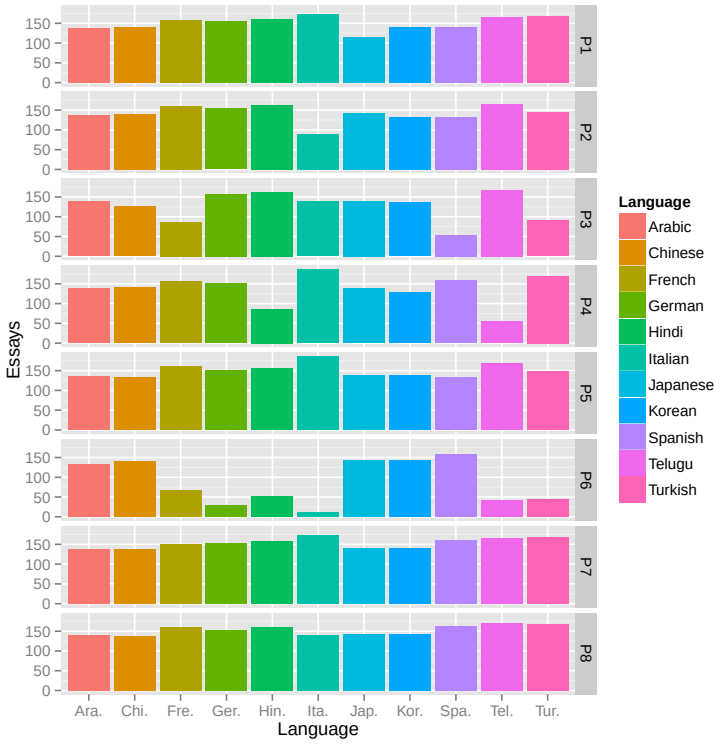

***Figure* 2 Number of essays per language per prompt.**

2. If two scores on an item differ by more than 1 point, a third human rater will read the item and give a score, resulting in three scores. If these three scores are now adjacent to one another, then average the three scores. For example, if the two scores from Rater 1 and Rater 2 are 3 and 5, and the essay goes to Rater 3, who then assigns a score of 4, then the scores are now 3, 5, and 4, and they are adjacent to one another (with no more than 1 point in difference: 3, 4, 5). Therefore, the average score (4) is the final score on this item.

3. If Rater 3 is involved and the three scores on an item contains an outlier, then average the two scores that are close together (adjacent) for the final score on this item. For example, when Rater 1 and Rater 2 give 2 and 4 to an item, and Rater 3 comes in and assigns a 5 for this item, then the three scores are 2, 4, and 5. Score 2 is deemed as an outlier and discarded. Scores 4 and 5 are averaged and 4.5 is the final score.

4. In the rare instance when the three scores are not adjacent to one another (1, 3, and 5, which is the only possible case), a fourth adjudicated score is the final score.

5. If any rater assigns the score of 0, the response goes to adjudication and the adjudicated score, which may be 0, 1, 2, 3, 4, or 5, is final. For example, when 0 is given on one item and then goes to adjudication, and the adjudicated score is 5, this item's score is then 5.

When collapsing the combined scores into the 3-point scale, *low* is for essays scoring between 1.0 and 2.0, *medium* is for 2.5 to 3.5, and *high* is for 4.0 to 5.0.[9] Table 2 and Figure 3 show the number of essays written by students belonging to each L1 group that received a low, medium, or high score. There are substantial score differences by L1 group, which reflect the characteristics of the sample from which these essays were drawn. It could be argued that a stratified sampling of essays (in which essays were selected to ensure that the score distribution was approximately the same for each language group) would be superior for the purpose of the NLI task, as the identification of L1s would not then be conflated with the task of gauging the quality of an essay. However, it was decided not to sample in this manner, as differences in essay quality (and ultimately, English writing proficiency) can also be taken as important and valid characteristics that distinguish the writing styles of L1 groups.

Table 2
Number of Essays per Language per Score Level

Language	Low	Medium	High
Arabic	296	605	199
Chinese	98	727	275
French	63	577	460
German	15	412	673
Hindi	29	429	642
Italian	164	623	313
Japanese	233	679	188
Korean	169	678	253
Spanish	79	563	458
Telugu	94	659	347
Turkish	90	616	394
Total	1,330	6,568	4,202

2.4 Essay Length

Because the corpus includes both low-scoring essays (which are frequently short) and high-scoring essays (which are frequently long), the length of TOEFL11 essays is quite variable,

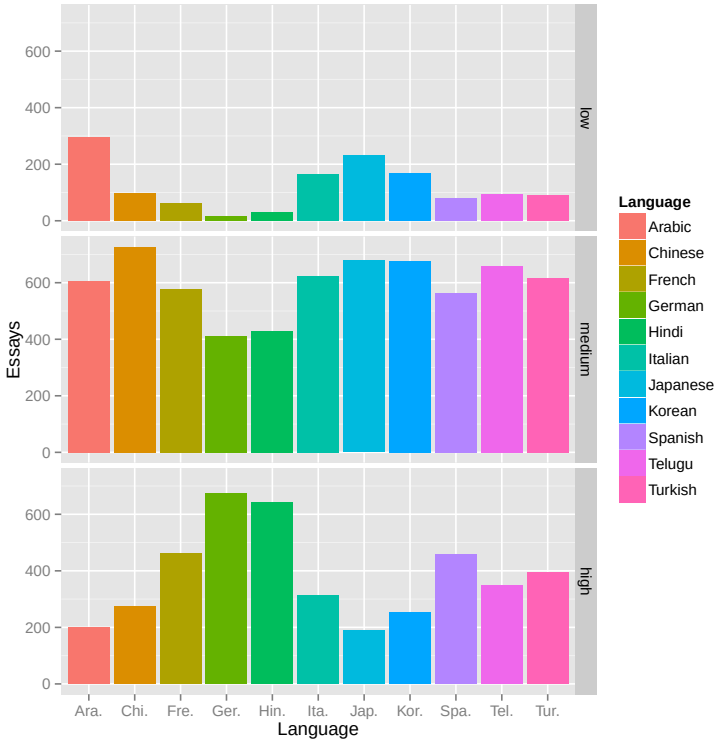

***Figure* 3 Number of essays per language per score level.**

ranging from two words to 876 words. That said, most essays were in the middle of the length distribution (see Figure 4). The length distributions for each prompt are shown in Figure 5 and Table 3, and the length distributions for each score level are shown in Table 4.

2.5 Language Families

The language family distribution for TOEFL11 strikes a balance between having many language families and many languages per family. As depicted in Figure 6, the language families present in TOEFL11 are Romance (French, Italian, Spanish), Germanic (German), Indo-Iranian (Hindi), Altaic (Japanese, Korean, Turkish),[10] Sino-Tibetan (Chinese), Afro-Asiatic (Arabic), and Dravidian (Telugu). The taxonomy of languages is deeply hierarchical, so it is worth noting that the Romance, Germanic, and Indo-Iranian languages are all members of the larger Indo-European group of languages. Altaic, Sino-Tibetan, Afro-Asiatic, and Dravidian are all at the top level of the language family taxonomy, so the other languages have far fewer common features than the Indo-European languages.

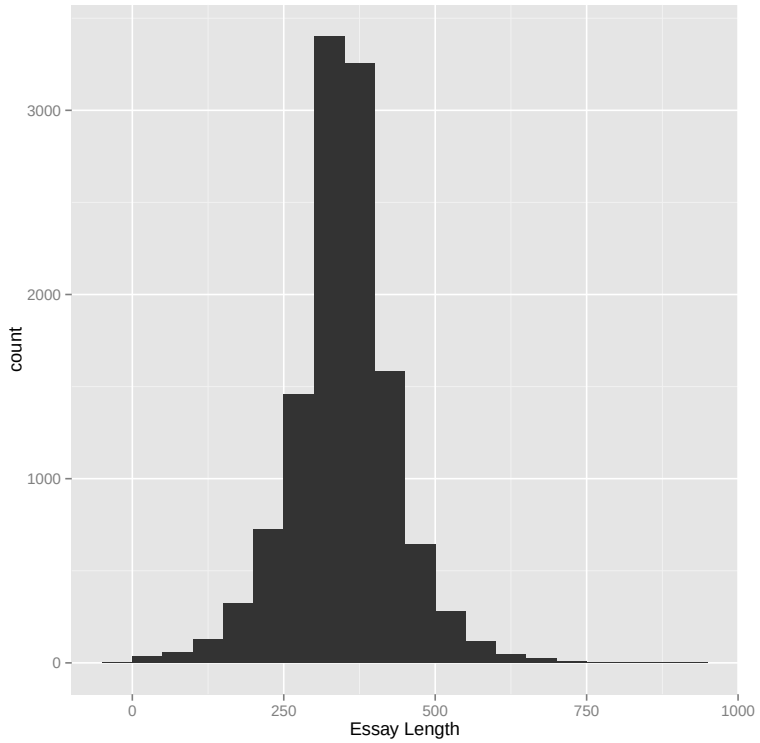

***Figure* 4 Histogram of essay lengths for all essays (bin width = 50 words).**

Table 3
Average Length of Essays (in Words) per Prompt per Language

Language	P1	P2	P3	P4	P5	P6	P7	P8
Arabic	320	327	282	312	317	318	295	306
Chinese	357	373	364	336	375	387	359	347
French	353	362	344	339	388	366	345	342
German	367	394	377	376	386	395	369	364
Hindi	379	383	407	373	413	370	362	377
Italian	306	329	334	324	324	352	331	325
Japanese	297	315	294	290	306	330	318	334
Korean	313	367	332	316	345	362	331	328
Spanish	355	385	352	358	357	385	341	355
Telugu	351	358	367	374	389	374	344	348
Turkish	347	360	337	364	367	357	330	346

The distribution of L1s in TOEFL11 is not typologically exhaustive and may therefore not exhibit all varieties of linguistic interference or other characteristics of writing that may be

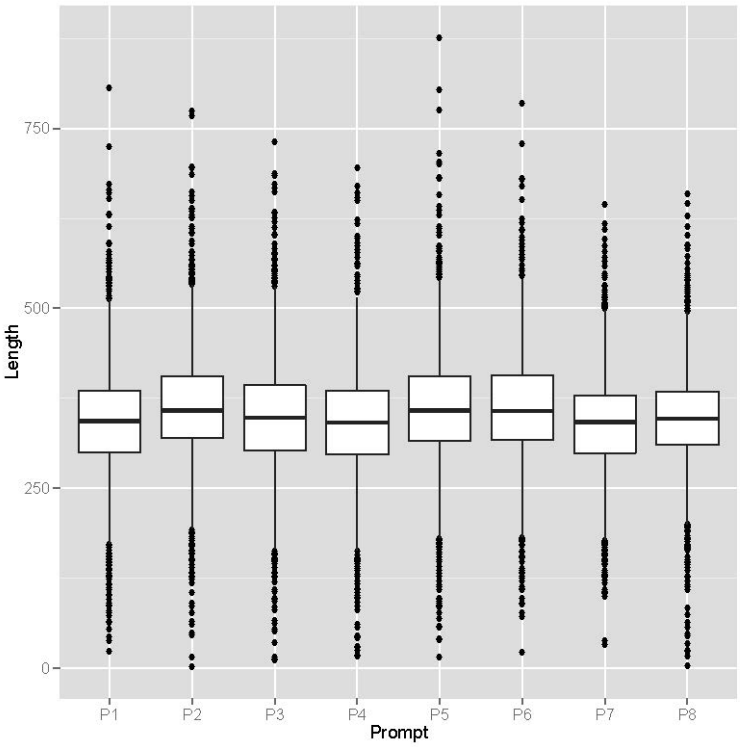

***Figure* 5 Distribution of essay lengths (in words) by prompt.**

conditioned on L1. However, because the sample is drawn from an operational test used for college admissions purposes, it is at least representative of the sample of languages spoken most frequently by non-native English speakers wishing to study abroad at a college where English is the language of instruction.

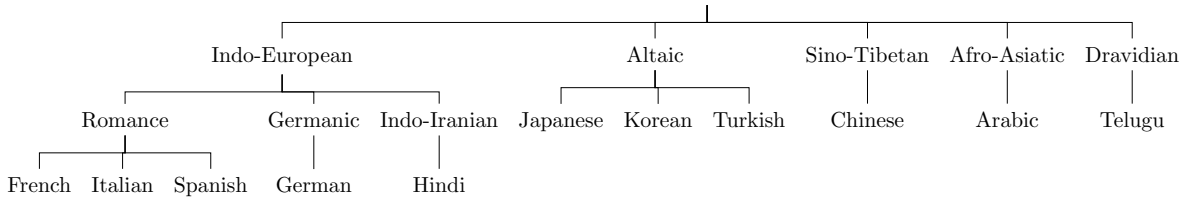

***Figure* 6 Language families in corpus.**

Table 4
Essay Length (in Words) Statistics for Each Score Level

	All	Low	Medium	High
Shortest	2	2	109	210
Longest	876	609	660	876
Mean	348	229	339	401
SD	85	86	61	72
# essays	12,100	1,330	6,568	4,202

3 Conclusion

The TOEFL11 dataset is the largest publicly available corpus of English written by non-native writers that is well-balanced for topic across L1s. It also is the first such corpus that is annotated for score level by highly trained human raters. Making it widely available will be a boon for researchers working on grammatical error detection and correction, NLI, and automatic essay scoring.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/1492

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., ... & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000* (Research Memorandum No. RM-04-03). Princeton, NJ: Educational Testing Service.

Comrie, B. (Ed.). (1990). *The world's major languages.* New York, NY: Oxford University Press.

Dahlmeier, D., & Ng, H. T. (2011, June). Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies* (pp. 915–923). Stroudsburg, PA: Association for Computational Linguistics.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D., & Vanderwende, L. (2008). Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the international joint conference on natural language processing (IJCLNP*, pp. 449–456). Hyderabad, India.

Granger, S., Dagneaux, E., & Meunier, F. (2009). *The International Corpus of learner English: Handbook and CD-ROM* (Version 2). Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Koppel, M., Schler, J., & Zigdon, K. (2005). Automatically determining an anonymous author's native language. In P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.-Y. Wang, H. Chen, & R. Merkle (Eds.), *Lecture notes in computer science: Vol. 3495. Intelligence and security informatics* (pp. 209–217). Berlin, Germany: Springer-Verlag.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated grammatical error detection for language learners.* San Rafael, CA: Morgan Claypool.

Rozovskaya, A., & Roth, D. (2010, June). Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 154–162). Stroudsburg, PA: Association for Computational Linguistics.

Shermis, M., & Hamner, B. (2012, April). *Contrasting state-of-the-art automated scoring of essays: Analysis.* Paper presented at the annual meeting of the National Council on

Measurement in Education. Vancouver, BC, Canada.

Swanson, B., & Charniak, E. (2012, July). Native language detection with tree substitution grammars. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics* (Vol. 2; pp. 193–197). Stroudsburg, PA: Association for Computational Linguistics.

Tetreault, J., Blanchard, D., & Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 48–57). Stroudsburg, PA: Association for Computational Linguistics.

Tetreault, J., Blanchard, D., Cahill, A., & Chodorow, M. (2012, December). Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)* (pp. 2585–2602). Stroudsburg, PA: Association for Computational Linguistics.

Tetreault, J., Burstein, J., & Leacock, C. (Eds.). (2012). *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications.* Stroudsburg, PA: Association for Computational Linguistics.

Wong, S.-M. J., & Dras, M. (2011, July). Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1600–1610). Stroudsburg, PA: Association for Computational Linguistics.

Wong, S.-M. J., Dras, M., & Johnson, M. (2011, December). Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 115–124). Stroudsburg, PA: Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011, June). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189). Stroudsburg, PA: Association for Computational Linguistics. Association for Computational Linguistics.

Notes

[1]Since the TOEFL11 corpus was first described here, it has been used as the basis for the first shared task in native language identification (NLI), which successfully took place at the 2013 NAACL workshop on Innovative Use of NLP for Building Educational Applications. For a full description of the shared task and some more recent developments in the field, please see Tetreault, Blanchard and Cahill (2013).

[2]See http://sigslate.org

[3]See http://aacl.sdsu.edu/call_for_papers.html

[4]See http://kaggle.com

[5]See http://langbank.engl.polyu.edu.hk/corpus/clec.html

[6]See http://r2m.nus.edu.sg/cos/o.x?c=/r2m/license_product&ptid=5730&func=viewProd&pid=28

[7]See http://nlp.comp.nus.edu.sg/conll13st/

[8]See http://www.ets.org/Media/Tests/TOEFL/pdf/Writing/Rubrics.pdf

[9]Essays with a combined score of 0 are invalid responses and were not included in TOEFL11.

[10]Note that the existence of an Altaic family encompassing all of these languages is a matter of some controversy (see Comrie, 1990).